Journal of Data Science 19 (3), 365–389
July 2021

DOI: 10.6339/21-JDS1009
Statistical Data Science

BayesSMILES: Bayesian Segmentation Modeling for Longitudinal Epidemiological Studies

Shuang Jiang[1,2], Quan Zhou[3], Xiaowei Zhan[2,*], and Qiwei Li[4,*]

[1]*Department of Statistical Science, Southern Methodist University, Dallas, TX 75205, USA*
[2]*Quantitative Biomedical Research Center, Department of Population and Data Sciences, The University of Texas Southwestern Medical Center, Dallas, TX 75390, USA*
[3]*Department of Statistics, Texas A&M University, College Station, TX 77843, USA*
[4]*Department of Mathematical Sciences, The University of Texas at Dallas, Richardson, TX 75080, USA*

Abstract

The coronavirus disease of 2019 (COVID-19) is a pandemic. To characterize its disease transmissibility, we propose a Bayesian change point detection model using daily actively infectious cases. Our model builds on a Bayesian Poisson segmented regression model that 1) capture the epidemiological dynamics under the changing conditions caused by external or internal factors; 2) provide uncertainty estimates of both the number and locations of change points; and 3) has the potential to adjust for any time-varying covariate effects. Our model can be used to evaluate public health interventions, identify latent events associated with spreading rates, and yield better short-term forecasts.

Keywords *Bayesian hierarchical modeling; multiple change-point detection; Poisson segmented regression; stochastic SIR model*

1 Introduction

A newly identified coronavirus, SARS-CoV-2, is a lethal virus for humans. It has caused a worldwide pandemic for the disease known as COVID-19. As reported by the Johns Hopkins University Center for Systems Science and Engineering (JHU-CSSE), the COVID-19 pandemic has spread to 188 countries and territories, with more than 14 million confirmed cases by the end of July 2020. The extremely rapid spreading of the disease and the increasing burden on healthcare systems have become major public health problems. In response to the public health demand to "flatten the curve" (Akiyama et al., 2020), both federal and local governments in the United States (U.S.) have enforced a wide range of public health measures, such as border shutdowns, travel restrictions, and quarantine.

As a consequence, the importance of understanding the COVID-19 dynamics is steadily increasing in the contemporary world. In epidemiology, the basic reproduction number, denoted by $\mathcal{R}_0$, is commonly used to evaluate the transmissibility of an infectious disease like COVID-19. $\mathcal{R}_0$ is interpreted as the expected number of secondary cases produced by a typical case in a closed population. During the outbreak of an epidemic, $\mathcal{R}_0$ can be affected by intervention strategies. For example, social measures that decrease the contact rate between individuals would control $\mathcal{R}_0$. Isolating or treating the infected cases could lower the $\mathcal{R}_0$ value as well. Another concept in the epidemic theory is the effective reproduction number $\mathcal{R}_t$, which describes the number of people who can be infected by an individual at any specific time t in a population. $\mathcal{R}_t$

*Corresponding authors. Email: Xiaowei.Zhan@UTSouthwestern.edu or Qiwei.Li@UTDallas.edu.

Received October 7, 2020; Accepted February 4, 2021

is time-specific since it accounts for the varying proportions of the population that are immune to the disease over time. There are many recent studies implementing the SIR model (Kermack and McKendrick, 1927) or its modified version to analyze COVID-19 transmissibility in terms of $\mathcal{R}_0$ or $\mathcal{R}_t$ (see e.g. Chen et al., 2020; Alvarez et al., 2020; Kantner and Koprucki, 2020; Gostic et al., 2020; Cooper et al., 2020). Furthermore, several studies have incorporated the information on social measures to understand the COVID-19 dynamics all over the world. For instance, Dehning et al. (2020) combined the SIR model with Bayesian inference to study the time-varying spreading rate of COVID-19 in Germany. Wang et al. (2020) extended the SIR model by considering the quarantine protocols with a focus on understanding the time-course dynamics of COVID-19 in Hubei, China. Giordano et al. (2020) enriched the SIR model with additional compartments to account for under-diagnosis, which could explain the gap between the actual infection dynamics and perception of COVID-19 outbreak in Italy. Because of the heterogeneity in susceptibility and social dynamics, COVID-19 transmissibility differs among locations and changes over time. U.S. local governments have implemented different interventions since mid-March to combat the spread of COVID-19. Therefore, the basic reproduction numbers should spatiotemporally vary.

The basic reproduction number of an epidemic event is changing due to societal and political actions. Effective social measures such as business closures and stay-at-home orders could help lower the transmission rate and thus induce changes in $\mathcal{R}_0$. By studying the variations in $\mathcal{R}_0$ over time, we can evaluate the dynamic transmissibility of infectious diseases like COVID-19. For instance, during the outbreak of severe acute respiratory syndrome (SARS) in China around 2003, it was reported an $\mathcal{R}_0 \approx 3.0$ for the onset stage of SARS in Hong Kong (Riley et al., 2003; Lloyd-Smith et al., 2003). Later on, it dropped to about 1.1 due to stringent control measures (Chowell et al., 2004). Decreases in $\mathcal{R}_0$ captured the evolution of SARS transmission dynamics under the approach of efficient diagnosis coupled with patient isolation measures. A recent study in Germany (Dehning et al., 2020) estimated the variations in COVID-19 transmission rates for the four pre-labeled phases partitioned by three time points corresponding to the three major government interventions. Meanwhile, Wang et al. (2020) extended the standard SIR model by introducing a transmission rate modifier, which takes different pre-specified decay functions under different macro or micro quarantine measures over time. These studies have enabled public health workers to analyze and evaluate the time-course dynamics of COVID-19 and motivated us to develop a method that can automatically detect the important transitioning time points that occurred during the outbreak of an epidemic, while characterizing the transmission dynamics.

We propose a method named BayesSMILES, which is short for Bayesian Segmentation ModelIng for Longitudinal Epidemiological Studies, to study the dynamics of COVID-19 transmissibility and to evaluate the effectiveness of mitigation interventions. BayesSMILES adopts a Bayesian Poisson segmented regression model to detect multiple change points based on the daily infectious COVID-19 cases. This novel model can 1) capture the epidemiological dynamics under the changing conditions caused by external or internal factors; 2) quantify the uncertainty in both the number and locations of change points; and 3) adjust any relevant explanatory time-varying covariates that may affect the infectious case numbers. In addition, BayesSMILES integrates the change point information to quantify the COVID-19 transmissibility by estimating the basic reproduction numbers in different segments. We demonstrate that our approach can improve the accuracy of the change point detection compared with a widely used change point search method on the simulated data. Applying the proposed BayesSMILES to the U.S. state-level COVID-19 daily report data, we find that the detected change points correlate well with the timing of publicly announced interventions. We also demonstrate that a stochastic SIR

model incorporating change point information can provide a better short-term forecast. In all, BayesSMILES enables us to understand the disease dynamics of COVID-19 and provides useful insights for the anticipation and control of current and future pandemics.

The rest of the paper is organized as follows. We review the traditional susceptible-infectious-recovered (SIR) model in Section 2. In Section 3, we describe the framework of BayesSMILES. The Markov chain Monte Carlo (MCMC) algorithm and posterior inference procedures are described in Section 4. We provide a comprehensive simulation study to illustrate the performance of the proposed method against a competing approach in Section 5. In Section 6, we analyze the COVID-19 data for U.S. states using the proposed BayesSMILES. Finally, we conclude with remarks in Section 7 and provide information about implementation in Supplement A.

2 Review of the SIR Model

The susceptible-infected-removed (SIR) model was developed to simplify the mathematical modeling of human-to-human infectious diseases by Kermack and McKendrick (1927). It is a fundamental compartmental model in epidemiology. At any given time, each individual in a closed population with size N is assigned to three distinctive compartments with labels: susceptible (S), infectious (I), or removed (R, being either recovered or deceased). The standard SIR model in continuous time that models the flow of people from S to I and then from I to R is described by the following set of nonlinear ordinary differential equations (ODEs):

$$\begin{cases} \frac{dS(t)}{dt} &= -\beta N^{-1} S(t) I(t) \\ \frac{dI(t)}{dt} &= \beta N^{-1} S(t) I(t) - \gamma I(t) \\ \frac{dR(t)}{dt} &= \gamma I(t) \end{cases} \tag{1}$$

for $t > 0$, subjecting to $S(t) + I(t) + R(t) = N$. Here $\beta > 0$ is the diseases transmission rate and $\gamma > 0$ is the removal (recovery or death) rate. Conceptually, susceptible individuals become infectious ($S \to I$) and then are ultimately removed from the possibility of spreading the disease ($I \to R$) due to death or recovery with immunity against reinfection.

The rationale behind the first equation in (1) is that the number of new infections during an infinitesimal amount of time, $-dS(t)/dt$, is equal to the number of susceptible people, $S(t)$, times the product of the contact rate, $I(t)/N$, and the disease transmission rate β. The third equation in (1) reflects that the infectious individuals leave the infectious population per unit of time, $dI(t)/dt$, as a rate of $\gamma I(t)$. The second equation in (1) follows from the first and third ones as a result of $dS(t)/dt + dI(t)/dt + dR(t)/dt = 0$. Assuming that only a small fraction of the population is infected or removed in the onset phase of an epidemic, we have $S(t)/N \approx 1$ and thus the second equation reduces to $dI(t)/dt = (\beta - \gamma)I(t)$, revealing that the infectious population is growing if and only if $\beta > \gamma$. As the expected lifetime of an infected case is given by γ^{-1}, the ratio β/γ is the average number of new infectious cases directly produced by an infected case in a completely susceptible population. Since it is a good indicator of the transmissibility of an infectious disease, the epidemiologists name it the *basic reproduction number* $\mathcal{R}_0 = \beta/\gamma$ in the context of a standard SIR model, or the *effective reproduction number* $\mathcal{R}_t = \beta_t/\gamma_t$ in the context of a time-variant SIR model, where β and γ are replaced by $\beta(t)$ and $\gamma(t)$ in (1).

The standard SIR model is appealing due to its simplicity. It can be extended in many different ways to better characterize the disease, such as considering vital dynamics, adding more compartments, and allowing more possible transitions between compartments. For instance, the susceptible-exposed-infectious-recovered (SEIR) model includes an additional compartment

accounting for the incubation period. The susceptible-infectious-recovered-susceptible (SIRS) model allows recovered individuals to return to a susceptible state. For a comprehensive summary, see Bailey et al. (1975), Becker and Britton (1999), Allen (2008), or Andersson and Britton (2012). It is worth noting that some modified SIR models are currently being used to model the COVID-19 outbreak under under-reporting scenarios (see e.g. Flaxman et al., 2020; Riou et al., 2020; Wang et al., 2020).

3 The Proposed BayesSMILES Method

3.1 Data Notations

During a pandemic such as COVID-19, the most accessible and complete data are the daily reported numbers on confirmed cases and deaths. Suppose N is the total population size in a given region. Let $\boldsymbol{C} = (C_1, \ldots, C_T)$ and $\boldsymbol{D} = (D_1, \ldots, D_T)$ be the sequences of cumulative confirmed case and death numbers observed at T successive equally spaced points in time (e.g. day), where C_t and $D_t \in \mathbb{N}$ for $t = 1, \ldots, T$. For a region for which recovery cases are closely monitored day by day, we use $\boldsymbol{E} = (E_1, \ldots, E_T)$ to denote the sequence of cumulative recovery case numbers. Thus, due to the compositional nature of the basic SIR model, the three trajectories can be constructed as $\boldsymbol{S} = (S_1, \ldots, S_T)$ with $S_t = N - C_t$, $\boldsymbol{R} = (R_1, \ldots, R_T)$ with $R_t = D_t + E_t$, and $\boldsymbol{I} = (I_1, \ldots, I_T)$ with $I_t = N - S_t - R_t = C_t - D_t - E_t$. For a region for which recovery cases do not exist or are under-reported, we consider both $\boldsymbol{R}$ and $\boldsymbol{I}$ as missing data and reconstruct these two sequences according to the last equation of (1) with a pre-specified constant removal rate γ. Specifically we set $I_1 = C_1$ and $R_1 = 0$, and generate $R_t = R_{t-1} + \lceil \gamma I_{t-1} \rceil$ and $I_t = I_{t-1} + (C_t - C_{t-1}) - (R_t - R_{t-1})$ from $t = 2$ to T sequentially, where $\lceil \cdot \rceil \colon \mathbb{R}^+ \to \mathbb{N}$ denotes the ceiling function. For the choice of γ, we suggest estimating its value from publicly available reports in some region where confirmed, death, and recovery cases are all well-documented, or from prior epidemic studies due to the same under-reporting issue in actual data. Lastly, given a vector $\boldsymbol{Y} = (Y_1, \ldots, Y_T)$ and some initial value Y_0 (for example, $\boldsymbol{Y}$ can be $\boldsymbol{C}$, $\boldsymbol{D}$, $\boldsymbol{E}$, $\boldsymbol{S}$, $\boldsymbol{I}$ or $\boldsymbol{R}$), we use $\dot{\boldsymbol{Y}} = (\dot{Y}_1, \ldots, \dot{Y}_T)$ to denote the lag one difference of $\boldsymbol{Y}$, where $\dot{Y}_t = Y_t - Y_{t-1}$ for $t = 1, \ldots, T$; that is, $\dot{Y}_t$ is the difference between two adjacent observations. Tables S1 and S2 in the Supplementary Material summarize the data notations as well as the key notations of models introduced in Sections 3.3 and 3.4, respectively.

3.2 Modeling Epidemic Dynamics via a Modified Stochastic SIR Model

An SIR model has three trajectories, one for each compartment. The compositional nature of the three trajectories, i.e. $S(t) + I(t) + R(t) = N$, implies that we need only two of them to solve the ODEs as shown in (1). As mentioned previously, assuming $S(t) \approx N$ for all t results in $dI(t)/dt = (\beta - \gamma)I(t)$ and further leads to an exact solution: $I(t) = I(0)\exp\{(\beta - \gamma)t\}$. For modeling daily reported actively infectious data $\boldsymbol{I}$, we utilize its discrete-time version,

$$I_t \quad = \quad I_0 \exp\{(\beta_t - \gamma)t\} \tag{2}$$

with a time-varying rate β_t to account for the transmissibility changes of the disease. For simplicity's sake, we assume a constant removal rate γ. Based on (2), we introduce a probabilistic model, which approximately mimics the dynamics of the deterministic standard SIR model as shown in (1) during the onset phase of a pandemic. Specifically, we suppose the infectious population

size at time t is sampled from a Poisson model,

$$I_t|\alpha_t \quad \sim \quad \text{Poi}(N\alpha_t), \quad t = 1, \ldots, T, \tag{3}$$

where $\alpha_t = I_0 \exp\{(\beta_t - \gamma)t\}/N$ is a redefined time-varying transmissibility parameter that depends on the initial infectious population size I_0, the disease transmission rate β_t, the removal rate γ, and any latent factors (e.g. public health intervention, social behavior, virus mutation, healthcare quality, etc.) that may affect the disease transmissibility. This model automatically accounts for measurement errors and uncertainties associated with the counts. Note that (3) can be generalized to a negative binomial (NB) model, i.e. $I_t|\alpha_t \sim \text{NB}(N\alpha_t, \phi_I)$ if needed, where ϕ_I is a dispersion parameter aiming to account for over-dispersion that might be observed in $\boldsymbol{I}$. Here we use $\text{NB}(\mu, \phi)$, $\mu, \phi > 0$ to denote an NB distribution with mean μ and variance $\mu + \mu^2/\phi$.

3.3 Detecting Change Points via a Poisson Segmented Regression Model

Our change point detection builds upon the above modified stochastic time-variant SIR model with stationary transmissibility α_t in a certain segment. We assume that β_t only changes at certain time points. Identifying those change points is of significant importance, as it not only enables us to characterize the temporal dynamics of the pandemic but also helps policymakers evaluate the effectiveness of the past and ongoing mitigation and intervention strategies.

In this paper, the change points are defined as those discrete time points that significantly alter the disease transmission rate β_t between two adjacent segments, given a constant removal rate γ across all time points. We introduce a latent binary vector $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_T)$, $\zeta_t \in \{0, 1\}$, with $\zeta_t = 1$ if time point t is a change point and $\zeta_j = 0$ otherwise. We set $\zeta_1 = 1$ by default, interpreting the first time point as the "zeroth change point." Those points with $\zeta_j = 0$ can be partitioned into segments bounded by two adjacent change points. Thus, we use another vector $\boldsymbol{z} = (z_1, \ldots, z_T)$, $z_t \in \{1, \ldots, K\}$ to reparameterize $\boldsymbol{\zeta}$, where we define $z_t = \sum_{u=1}^{t} \zeta_u$. Thus, $z_t = k$ indicates that time point t is in segment k, that is, between the $(k-1)$ and k-th change points. The total number of change points excluding the first time point is $K - 1$. Note that $\boldsymbol{\zeta}$ is the lag one difference of $\boldsymbol{z}$, i.e. $\zeta_t = z_t - z_{t-1}$ with $\zeta_1 = 1$. Figure 1 shows the representations of $\boldsymbol{\zeta}$ and $\boldsymbol{z}$ for a simulated time-series dataset ($T = 10$) with two change points.

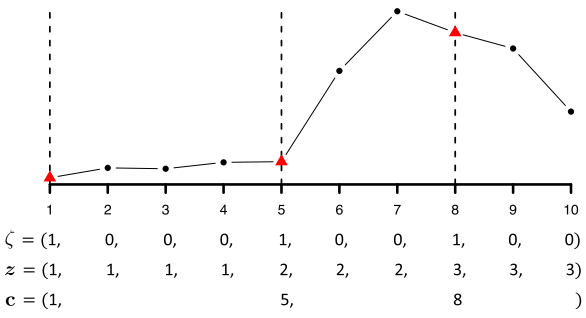


Figure 1: An example of time-series data ($T = 10$) with two change points ($K = 3$) and its associated parameterizations in terms of $\boldsymbol{\zeta}$ and $\boldsymbol{z}$, respectively. Red triangles are change points, while black circles are not. Note that the first time point is treated as the "zeroth change point."

To infer $\boldsymbol{\zeta}$ or $\boldsymbol{z}$ given the sequence of infectious population size $\boldsymbol{I}$, we adopt a Poisson segmented regression framework, which can be written as,

$$\begin{aligned} I_t|\alpha_t &\sim \text{Poi}(N\alpha_t), \quad t = 1, \ldots, T \\ \tilde{\alpha}_t|\boldsymbol{b}_k, z_t = k &= \boldsymbol{x}_t\boldsymbol{b}_k + \epsilon_t, \end{aligned} \tag{4}$$

where $\tilde{\alpha}_t = \log \alpha_t$, $\boldsymbol{x}_t = (1, t, x_{t,1}, \ldots, x_{t,p-2})$ is a p-dimensional row vector of covariates that includes a scalar of one for the intercept, time t, and $p-2$ explanatory variables observed at time t if applicable. Those explanatory variables could contain the number of tests, weather information, mobility report, or other necessary and accessible time-varying measures important to adjust for during a longitudinal epidemiological study. The vector $\boldsymbol{b}_k = (b_{1,k}, \ldots, b_{p,k})^\top$ is a p-dimensional column vector of segment-specified coefficients that includes an intercept representing the proportion of infectious people at logarithmic scale, i.e. $b_{1,k} = \log(I_0/N)$, in segment k, and a slope accounting for the time-varying disease transmission rate, i.e. $b_{2,k} = \beta_t - \gamma$. Let $\boldsymbol{X}$ denote the design matrix, which combines all $\boldsymbol{x}_t$'s as rows and $\boldsymbol{B}$ denote the corresponding coefficient matrix, which combines all $\boldsymbol{b}_k$'s as columns. For simplicity's sake, we assume the process error $\epsilon_1, \ldots, \epsilon_T$ are independent and identically Gaussian distributed with zero mean and segment-specified variance, i.e. $\epsilon_t \sim \text{N}(0, \sigma_k^2)$. To ensure the identifiability of all model parameters, we try a set of considerably small values for σ_k^2's and employ a robust cross validation method called Pareto-smoothed importance sampling leave-one-out (PSIS-LOO) cross validation to determine the best choice (Vehtari et al., 2017).

Let $\boldsymbol{\alpha}_k$ be the sequence of all α_t's in segment k, i.e. $\boldsymbol{\alpha}_k = (\alpha_{c_k}, \ldots, \alpha_{c_k+n_k-1})^\top$, where we denote $c_k = \min\{t : z_t = k\}$ as the location of the $(k-1)$-th change point and $n_k = \sum_{t=1}^{T} \delta(z_t = k)$ as the number of time points in segment k with $\delta(\cdot)$ being the indicator function. We can rewrite the second equation in (4) as $\tilde{\boldsymbol{\alpha}}_k|\boldsymbol{b}_k \sim \text{MN}(\boldsymbol{X}_k\boldsymbol{b}_k, \sigma_k^2\mathcal{I}_{n_k})$, where $\mathcal{I}_{n_k}$ is an n_k-by-n_k identity matrix and $\boldsymbol{X}_k$ can be explicitly written as

$$\begin{pmatrix} 1 & t_{c_k} & x_{c_k,1} & \cdots & x_{c_k,p-2} \\ 1 & t_{c_k+1} & x_{c_k+1,1} & \cdots & x_{c_k+1,p-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{c_k+n_k-1} & x_{c_k+n_k-1,1} & \cdots & x_{c_k+n_k-1,p-2} \end{pmatrix}.$$

We assume a zero-mean multivariate normal distribution for $\boldsymbol{b}_k$, that is, $\boldsymbol{b}_k \sim \text{MN}(\boldsymbol{0}_p, \boldsymbol{H})$, where $\boldsymbol{0}_p$ is an p-by-1 all-zero column vector and $\boldsymbol{H} = \text{Diag}(h_0, \ldots, h_{p-1})$ is set to be a diagonal variance-covariance matrix. For a weakly informative setting, we recommend choosing a large value for each h_j. Through this prior specification, we are able to marginalize out the nuisance parameter $\boldsymbol{b}_k$ and obtain $\tilde{\boldsymbol{\alpha}}_k \sim \text{MN}\left(\boldsymbol{0}_{n_k}, \boldsymbol{X}_k\boldsymbol{H}\boldsymbol{X}_k^\top + \sigma_k^2\mathcal{I}_{n_k}\right)$. Consequently, we can write the collapsed model of (4) as

$$\begin{aligned} I_1, \ldots, I_T|\boldsymbol{\alpha} &\sim \prod_{t=1}^{T} \text{Poi}(N\alpha_t), \\ \tilde{\boldsymbol{\alpha}}|\boldsymbol{\zeta} &\sim \prod_{k=1}^{K} \text{MN}\left(\boldsymbol{0}_{n_k}, \boldsymbol{X}_k\boldsymbol{H}\boldsymbol{X}_k^\top + \sigma_k^2\mathcal{I}_{n_k}\right). \end{aligned} \tag{5}$$

To complete the model specification, we impose an independent Bernoulli prior on $\boldsymbol{\zeta}$ as $\boldsymbol{\zeta}|\omega \sim \prod_{t=2}^{T} \text{Bern}(\omega)$, where ω is interpreted as the probability of a time point being a change point *a priori*. We further relax this assumption by allowing $\omega \sim \text{Be}(a_\omega, b_\omega)$ to achieve a beta-Bernoulli

prior. In practice, we suggest a constraint of $a_\omega + b_\omega = 2$ for a vague hyperprior of ω (Tadesse et al., 2005). We make the prior probability of $\boldsymbol{\zeta}$ equal to zero if two adjacent time points are jointly selected as change points (i.e. a segment should consist of at least two time points).

3.4 Estimating Basic Reproduction Numbers via a Stochastic SIR Model

Given the multiple change points $\boldsymbol{\zeta}$, we estimate the basic reproduction number $\mathcal{R}_0 = \beta/\gamma$ for each segment via a stochastic version of the standard SIR model as shown in (1), which only needs the cumulative confirmed case numbers $\boldsymbol{C}$. This is because recovery data exist in only a few states in the U.S., which makes both model inference and predictions infeasible. This model considers both of the removed and actively infectious cases as missing data and mimics their relationship as in some compartmental models in epidemiology. Specifically, we assume the number of new removed cases at time t, i.e. $\dot{R}_t$, is sampled from a Poisson distribution with mean γI_{t-1}, that is, $\dot{R}_t \sim \text{Poi}(\gamma I_{t-1}) = \text{Poi}(\gamma(N - C_{t-1} - R_{t-1}))$, where γ should be pre-specified. Based on this simplification, we rewrite the discrete version of the first equation in (1) as,

$$(N - C_t) - (N - C_{t-1})|z_t = k \quad = \quad -\beta_k^*(N - C_{t-1})\frac{N - C_{t-1} - R_{t-1}}{N},$$

resulting in

$$\dot{C}_t|z_t = k \quad = \quad \beta_k^*(N - C_{t-1})\frac{N - C_{t-1} - R_{t-1}}{N},$$

where β_k^* is the common disease transmission rate for the all the time points in segment k.

We assume the new case number observed at time t, i.e. $\dot{C}_t$, is sampled from an NB model,

$$\dot{C}_t|z_t = k \quad \sim \quad \text{NB}\left(\beta_k^*(N - C_{t-1})\frac{N - C_{t-1} - R_{t-1}}{N}, \phi_k\right), \quad t = 2, \ldots, T, \tag{6}$$

as it automatically accounts for measurement errors and uncertainties associated with the counts. Following most epidemiological models, we assume this stochastic process is a Markov process, where the present state (at time t) depends only upon its previous state (at time $t-1$). The setting above builds upon the standard SIR model. It is worth noting that the oversimplified assumptions of the proposed stochastic SIR model, as well as the bias in data reporting, might undermine the reliability of the estimates on disease transmission rates β_k^*'s and their succeeding basic reproduction numbers $\mathcal{R}_{0k}$'s. However, they can still be used as a proxy to indicate the transmissibility dynamic of an infectious disease. We could consider additional compartments as seen in the susceptible-infectious (SIS) model, the susceptible-infectious-recovered-deceased (SIRD) model, the susceptible-exposed-infectious-removed (SEIR) model, and the susceptible-exposed-infectious-susceptible (SEIS) model (see a comprehensive summary in Bailey et al., 1975). The effects from the additional compartments could be incorporated by reparameterizing the mean function in the NB distribution, as shown in Equation (6), which is left as future work. For the prior distribution of the segment-specific dispersion parameter ϕ_k, we choose a gamma distribution, $\phi_k \sim \text{Ga}(a_\phi, b_\phi)$ for $k = 1, \ldots, K$. We recommend small values, such as $a_\phi = b_\phi = 0.001$, for a weakly informative setting. This model, on average, mimics the epidemic dynamics and is more flexible than those deterministic epidemiological models. For each segment k, we assume β_k^* comes from a gamma distribution with hyperparameters that makes both mean and variance of the transformed variable β_k^*/γ equal to 1.

4 Model Fitting

In this section, we describe the Markov chain Monte Carlo (MCMC) algorithms for posterior inference of the proposed BayesSMILES method, including the inferential strategy for identifying change points and estimating the basic reproduction numbers, respectively. See Section S4 in the Supplementary Material for details of our MCMC algorithm. Although it is feasible to use an established probabilistic programming language (PPL) such as Stan (Gelman et al., 2015) or PyMC3 (Salvatier et al., 2016) to fit the model, we prefer to derive our own MCMC algorithm to have more control over the implementation. For example, it makes us easy to adopt a technique (detailed in Section S4.3 in the Supplementary Material) to reduce the computational complexity.

4.1 MCMC Algorithms for Detecting Change Points

Our primary interest lies in identifying the change point locations via the vector $\boldsymbol{\zeta}$ based on the actively infectious cases $\boldsymbol{I}$. According to Section 3.3, the full data likelihood and the priors of the change point detection model are written as,

$$\begin{aligned} f(\boldsymbol{I}|\boldsymbol{\alpha}) &= \prod_{t=1}^{T} \mathrm{Poi}(I_t; N\alpha_t) \\ \pi(\tilde{\boldsymbol{\alpha}}|\boldsymbol{\zeta}) &= \prod_{k=1}^{K} \mathrm{MN}\left(\tilde{\boldsymbol{\alpha}}_k; \mathbf{0}_{n_k}, \boldsymbol{X}_k \boldsymbol{H} \boldsymbol{X}_k^\top + \sigma_k^2 \mathcal{I}_{n_k}\right) \\ \pi(\boldsymbol{\zeta}) &= \prod_{t=2}^{T} \text{Be-Bern}(\zeta_t; a_\omega, b_\omega). \end{aligned} \tag{7}$$

Thus, the full posterior distribution is $\pi(\boldsymbol{\alpha}, \boldsymbol{\zeta}|\boldsymbol{I}) \propto f(\boldsymbol{I}|\boldsymbol{\alpha}, \boldsymbol{\zeta})\pi(\boldsymbol{\alpha}, \boldsymbol{\zeta}) = f(\boldsymbol{I}|\boldsymbol{\alpha})\pi(\tilde{\boldsymbol{\alpha}}|\boldsymbol{\zeta})\pi(\boldsymbol{\zeta})$. Since there are no closed form expressions for the two conditionals $\pi(\boldsymbol{\alpha}|\boldsymbol{\zeta}, \boldsymbol{I})$ and $\pi(\boldsymbol{\zeta}|\boldsymbol{\alpha}, \boldsymbol{I})$, we use Metropolis-Hastings (MH) algorithms to sample from the two distributions. We also tried the gradient-based Metropolis-adjusted Langevin algorithm (Roberts and Rosenthal, 1998) and found no noticeable difference in the change-point inference.

4.2 MCMC Algorithms for Estimating Basic Reproduction Numbers

Once the change points are determined, we aim to estimate the basic reproduction numbers $\mathcal{R}_0$'s across different segments and quantify their uncertainties based on the cumulative confirmed cases $\boldsymbol{C}$ only. According to Section 3.4, the full data likelihood and the priors of the stochastic SIR model are written as,

$$\begin{aligned} f(\dot{\boldsymbol{C}}|\boldsymbol{\beta}^*, \boldsymbol{\phi}, \boldsymbol{R}) &= \prod_{k=1}^{K} \prod_{\{t:z_t=k\}} \mathrm{NB}\left(\dot{C}_t; \beta_k^*(N - C_{t-1})\frac{N - C_{t-1} - R_{t-1}}{N}, \phi_k\right) \\ \pi(\boldsymbol{\beta}^*) &= \prod_{k=1}^{K} \mathrm{Ga}(\beta_k^*; a_\beta, b_\beta) \\ \pi(\boldsymbol{\phi}) &= \prod_{k=1}^{K} \mathrm{Ga}(\phi_k; a_\phi, b_\phi), \end{aligned} \tag{8}$$

where $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_K^*)$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$, i.e. the collections of transmission and dispersion rates of all segments. For the hyperparameters, we set $a_\beta = 1$ and $b_\beta = 1/\gamma$ so that both of the expectation and variance of the basic reproduction number $\mathcal{R}_0 = \beta_k^*/\gamma$ are equal to one. With a pre-defined removal rate γ, we propose the following updates in each MCMC iterations.

4.3 Posterior Inference

We explore posterior inference for the parameters of interest by postprocessing the MCMC samples after burn-in iterations. We start by obtaining a point estimate of the change point indicator $\boldsymbol{\zeta}$ by analyzing its MCMC samples $\{\boldsymbol{\zeta}^{(u)}, \ldots, \boldsymbol{\zeta}^{(U)}\}$, where u indexes the MCMC iteration after burn-in. One way is to choose the $\boldsymbol{\zeta}$ corresponding to the *maximum-a-posteriori* (MAP),

$$\hat{\boldsymbol{\zeta}}^{\text{MAP}} = \underset{u}{\text{argmax}}\ \pi(\boldsymbol{\alpha}^{(u)}|\boldsymbol{\zeta}^{(u)})\pi(\boldsymbol{\zeta}^{(u)}).$$

The corresponding $\hat{\boldsymbol{z}}^{\text{MAP}}$ can be obtained by taking the cumulative sum of $\hat{\boldsymbol{\zeta}}^{\text{MAP}}$. An alternative estimate relies on the computation of posterior pairwise probability matrix (PPM), where the probability that time points t and t' are assigned into the same segment is estimated by $p_{tt'} \approx \sum_{u=1}^{U} \delta(z_t^{(u)} = z_{t'}^{(u)}|\cdot)$. This estimate utilizes the information from all MCMC samples and is thus more robust. After obtaining this T-by-T co-clustering matrix denoted by $\boldsymbol{P} = [p_{tt'}]_{T\times T}$, a point estimate of $\boldsymbol{z}$ can be approximated by minimizing the sum of squared deviations of its association matrix from the PPM, that is,

$$\hat{\boldsymbol{z}}^{\text{PPM}} = \underset{z}{\text{argmin}} \sum_{t<t'} (\delta(z_t = z_{t'}) - p_{tt'})^2 .$$

The corresponding $\hat{\boldsymbol{\zeta}}^{\text{PPM}}$ can be obtained by taking the difference between consecutive entries in $\hat{\boldsymbol{z}}^{\text{PPM}}$ and setting the first entry to one. To construct a "credible interval" for a change point, we utilize its local dependency structure from all MCMC samples of $\boldsymbol{\zeta}$ that belong to its neighbors. Due to the nature of the MCMC algorithm described in Section 4.1, if a time point t is selected as a change point, i.e. $\zeta_t = 1$, then its nearby time points must not be a change point. Thus, the correlation between the MCMC sample vectors $(\zeta_t^{(1)}, \ldots, \zeta_t^{(U)})$ and $(\zeta_{t\pm s}^{(1)}, \ldots, \zeta_{t\pm s}^{(U)})$ tends to be negative when s is small. We define the credible interval of a change point as the two ends of all its nearby consecutive time points, for which the MCMC samples of $\boldsymbol{\zeta}$ are significantly negatively correlated with that of the change point. This could be done via a one-sided Pearson correlation test with a pre-specified significant level, e.g. 0.05. Although quantifying uncertainties of change points is not rigorous, it performs very well in the simulation study and yields reasonable results in the real data analysis.

Once the change points are determined, an approximate Bayesian estimator of the disease transmission rate β_k^* for each segment k can be simply obtained by averaging over all of its MCMC samples, $\hat{\beta}_k = \sum_{u=1}^{U} \beta_k^{(u)}/U$. In addition, a quantile estimation or credible interval can be obtained. Lastly, we summarize the basic reproduction number in each segment k as $\hat{\mathcal{R}}_{0k} = \hat{\beta}_k/\gamma$.

4.4 Prediction

Conditional on the change point locations, we can predict the cumulative or new confirmed cases at any future time T_f by Monte Carlo simulation based on the information in the last segment K only. Specifically, from time $T+1$ to T_f, we sequentially generate

$$\dot{C}_t^{(u)} \quad \sim \quad \text{NB}\left(\beta_K^*(N - C_{t-1})\frac{N - C_{t-1} - R_{t-1}}{N}, \phi_K^{(u)}\right), \quad t = T+1, \ldots, T_f. \tag{9}$$

Then, both short and long-term forecasts can be made by summarizing the $(T_f - T)$-by-U matrix of MCMC samples. For instance, the predictive number of cumulative and new confirmed cases at time $T+1$, on average, are $\sum_{u=1}^{U} C_{T+1}^{(u)}/U$ and $\sum_{u=1}^{U} \dot{C}_{T+1}^{(u)}/U$, respectively.

5 Simulation

We used simulated data to evaluate the performance of our BayesSMILES method in terms of both change point detection and basic reproduction number estimation. It is shown that the proposed Bayesian framework outperforms an alternative change point detection method.

5.1 The Generative Model

The three trajectories $\boldsymbol{S}$, $\boldsymbol{I}$, and $\boldsymbol{R}$ with length $T = 120$ were generated in the following way. We first divided the $T = 120$ time points into $K = 4$ segments with the same length; that is, the true change points were $t = 31$, $t = 61$, and $t = 91$. To mimic the disease transmissibility dynamics across different segments, we chose segment-varying disease transmission rates β_k^* while fixing the removal rate $\gamma = 0.03$. Let $\boldsymbol{\mathcal{R}}_0$ be a K-vector where each entry gives the reproduction number of one segment, which can be computed by β_k^*/γ for $k = 1, \ldots, K$. We considered four scenarios of the set $(\beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*)$, corresponding to 1) $\boldsymbol{\mathcal{R}}_0 = (3.0,\ 1.2,\ 2.0,\ 0.8)$; 2) $\boldsymbol{\mathcal{R}}_0 = (3.0,\ 2.3,\ 1.5,\ 0.8)$; 3) $\boldsymbol{\mathcal{R}}_0 = (3.0,\ 1.8,\ 0.8,\ 1.6)$; 4) $\boldsymbol{\mathcal{R}}_0 = (3.0,\ 2.0,\ 1.1,\ 0.5)$. Then based on the stochastic version of the standard SIR model, we sampled S_t and R_t from negative binomial (NB) distributions, and obtained I_t, sequentially from $t = 1$ to T through

$$\begin{cases} S_t &= S_{t-1} - \mathrm{NB}(\beta_{z_{t-1}}^* N^{-1} S_{t-1} I_{t-1}, \phi_S) \\ R_t &= R_{t-1} + \mathrm{NB}(\gamma I_{t-1}, \phi_R) \\ I_t &= N - S_t - R_t \end{cases},$$

where $N = 1{,}000{,}000$, the initial $I_0 = 100$ and $R_0 = 0$, and the dispersion parameters $\phi_S = \phi_R = 10$. Note that the generative scheme was with an NB error structure, which was different from our model assumption based on a Poisson error structure. We repeated the above steps to generate 50 independent datasets for each setting of $\boldsymbol{\mathcal{R}}_0$. Figure 2 displays the temporal patterns of the simulated infectious counts $\boldsymbol{I}$ for the four scenarios.

5.2 Evaluation Criteria

To evaluate the change point detection, we may rely on either the binary change point indicator vector $\boldsymbol{\zeta}$ or the time point allocation vector $\boldsymbol{z}$. For the choice of $\boldsymbol{\zeta}$, a change point is considered to be correctly identified if its location is within a local window of the true position (Killick and Eckley, 2014). The selection of the window size is *ad hoc* and may bias the evaluation. In addition to that, since change points and non-change points are usually of very different sizes, most of the binary classification metrics are not suitable for model comparison here. Thus, we chose those metrics that quantify the agreement between the true and estimated allocation vectors, i.e. $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$. The two classic performance metrics for the analysis of clustering results are the adjusted Rand index (ARI) and mutual information (MI), proposed by Hubert and Arabie (1985) and Steuer et al. (2002), respectively. ARI is the corrected-for-chance version of the Rand index (Rand, 1971), as a similarity measure between two sample allocation vectors. Let $\mathrm{a} = \sum_{t>t'} \delta(z_t = z_{t'})\delta(\hat{z}_t = \hat{z}_{t'})$; $\mathrm{b} = \sum_{t>t'} \delta(z_t = z_{t'})\delta(\hat{z}_t \neq \hat{z}_{t'})$;

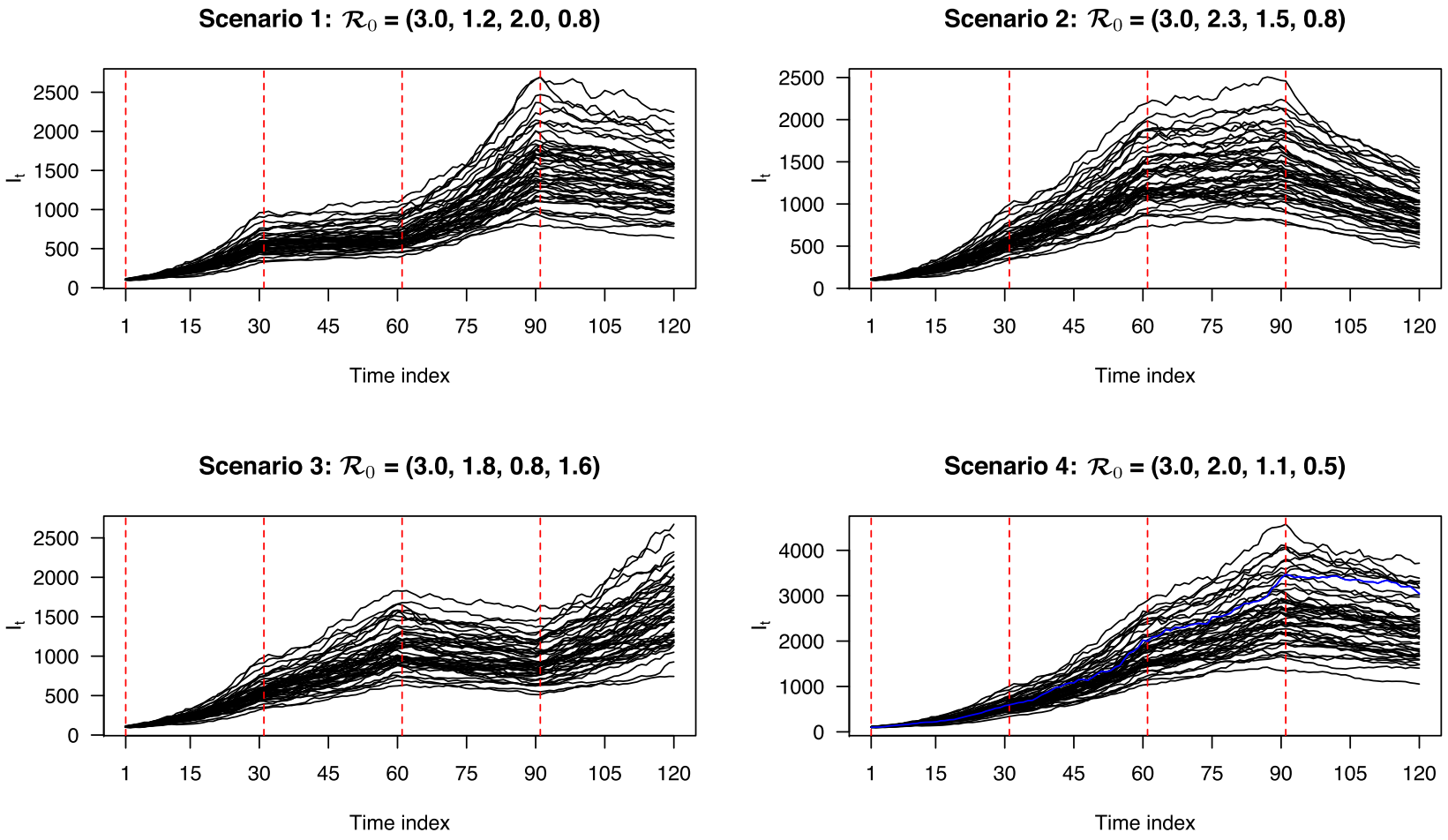


Figure 2: Simulation study: The simulated actively infectious data $\boldsymbol{I}$ under the four scenarios. Each curve represents a replicated sequence of $\boldsymbol{I}$ under a scenario. The red dashed lines mark the true change point locations. The blue curve under Scenario 4 was randomly chosen for evaluating the model fitting, of which results are shown in Figure 3.

$c = \sum_{t>t'} \delta(z_t \neq z_{t'})\delta(\hat{z}_t = \hat{z}_{t'})$; and $d = \sum_{t>t'} \delta(z_t \neq z_{t'})\delta(\hat{z}_t \neq \hat{z}_{t'})$ be the number of pairs of time points that are a) in the same segment in both of the true and estimated partitions; b) in different segments in the true partition but in the same segment of the estimated one; c) in the same segment of the true partition but in different segments in the estimated one; and d) in different segments in both of the true and estimated partitions. Then, the ARI can be computed as

$$\text{ARI}(\boldsymbol{z}, \hat{\boldsymbol{z}}) \quad = \quad \frac{\binom{T}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{\binom{T}{2}^2 - [(a+b)(a+c)+(c+d)(b+d)]}.$$

The ARI usually yields values between 0 and 1, although it can yield negative values (Santos and Embrechts, 2009). The larger the index, the more similarities between $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$, and thus the more accurately the method detects the actual times at which change points occurred. An alternative metric choice is MI, which measures the information about one variable that is shared by the other (Steuer et al., 2002). Let $m_{kk'} = \sum_{t=1}^{T} \delta(z_t = k)\delta(\hat{z}_t = k')$ be the number of time points shared between the k-th segment in the true $\boldsymbol{z}$ and the k'-th segment in the estimated one $\hat{\boldsymbol{z}}$. Then, MI can be computed as

$$\text{MI}(\boldsymbol{z}, \hat{\boldsymbol{z}}) \quad = \quad \sum_{k=1}^{K}\sum_{k'=1}^{\hat{K}} \frac{m_{kk'}}{T} \log \frac{m_{kk'}T}{n_k \hat{n}_{k'}},$$

where $\hat{K}$ is the number of segments and $\hat{n}_k$'s are the segment lengths for segment $1, \ldots, \hat{K}$ in $\hat{\boldsymbol{z}}$. It yields non-negative values. The larger the MI, the more accurate the partition result.

To quantify how well a method estimates the dynamic transmissibility across different segments, we used the root mean square error (RMSE) that measures the deviation between the true and estimated values of $\boldsymbol{\mathcal{R}}_0$ over all T time points:

$$\text{RMSE}(\boldsymbol{\mathcal{R}}_0, \hat{\boldsymbol{\mathcal{R}}}_0) \quad = \quad \sum_{t=1}^{T} (\mathcal{R}_{0z_t} - \hat{\mathcal{R}}_{0\hat{z}_t})/T.$$

A smaller value of RMSE indicates a more accurate estimation of $\mathcal{R}_0$'s.

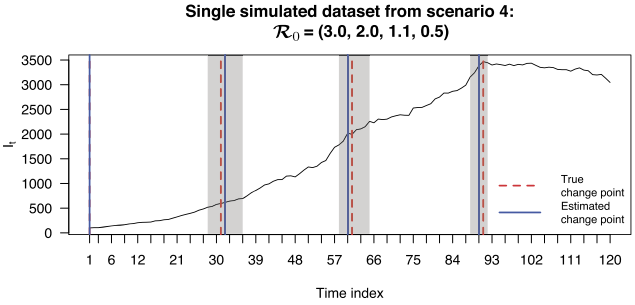
5.3 Results

As for the MCMC setting of change point detection, we set 40,000 MCMC iterations and discarded the first half as burn-in. We adopted the weakly informative setting by setting $a_\omega = 0.1$ and $b_\omega = 1.9$ in the Beta-Bernoulli prior for the change point indicator vector $\boldsymbol{\zeta}$. We set $\boldsymbol{H} = \text{Diag}(h_0, h_1)$ with $h_0 = 10{,}000$ and $h_1 = 10$ as the covariance matrix in the prior distribution of $\boldsymbol{b}_k$'s. Finally, we let σ_k^2 take ten equally spaced values ranging from 0.0001 to 0.01 at the logarithmic scale (base 10) in the PSIS-LOO cross validation. In fitting the stochastic SIR model, we set 100,000 MCMC iterations with the first half as burn-in. As suggested in Waqas et al. (2020), the value of removal rate γ could be estimated by $(T-1)^{-1}\sum_{t=2}^{T}(R_t - R_{t-1})/I_t$ for each simulated dataset. Then, we set $a_\beta = 1$ and $b_\beta = 1/\gamma$ so that both the prior expectation and prior variance of the basic reproduction number $\mathcal{R}_0 = \beta_k^*/\gamma$ are equal to 1.

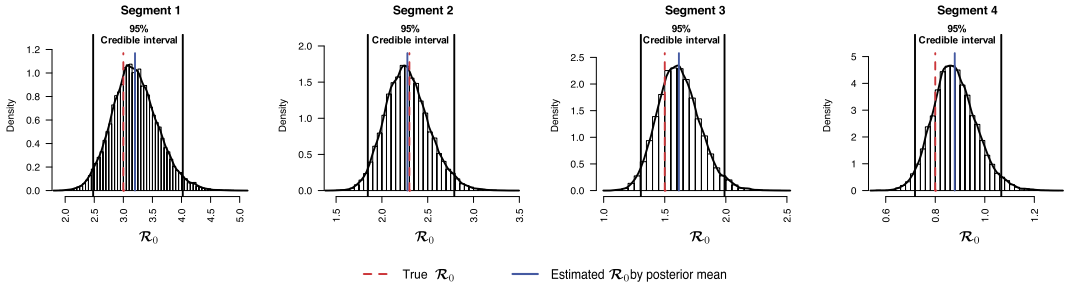
We first checked the performance of BayesSMILES on a single simulated dataset, which was randomly selected from the 50 replicates in Scenario 4 (marked as the blue line in Figure 2). Note

that we did the same for the remaining three scenarios, and the related results are summarized in Section S5 in the Supplementary Material. Figure 3(a) demonstrates the change point detection result based on the Poisson segmented regression model. The red dashed and the blue solid lines represent the true and the estimated change point locations, respectively, while the gray ribbons represent the 95% credible intervals for those identified change points. As we can see, BayesSMILES successfully detected the three true change points in general, as each of the 95% credible intervals covered the truth. The resulted values of ARI and MI were 0.93 and 1.28, respectively. Later on, the stochastic SIR model introduced in Section 3.4 was then fitted to quantify the disease transmissibility in each segment bounded by the identified change points. Figure 3(b) shows the posterior distributions of $\mathcal{R}_{0\hat{k}}$'s for $\hat{k} = 1, 2, 3, 4$ from their MCMC samples. The red dashed and blue solid lines pinpoint the true and posterior mean of $\mathcal{R}_{0\hat{k}}$'s, while the two black solid lines mark the boundary of their 95% credible intervals. Clearly, those true values were within their corresponding 95% credible intervals. The final RMSE for $\mathcal{R}_0$ estimation was 0.38.

To the best of our knowledge, there is no method like BayesSMILES that can detect latent change points while characterizing the transmission dynamics through an SIR model. Thus, in



(a) Change point detection performance

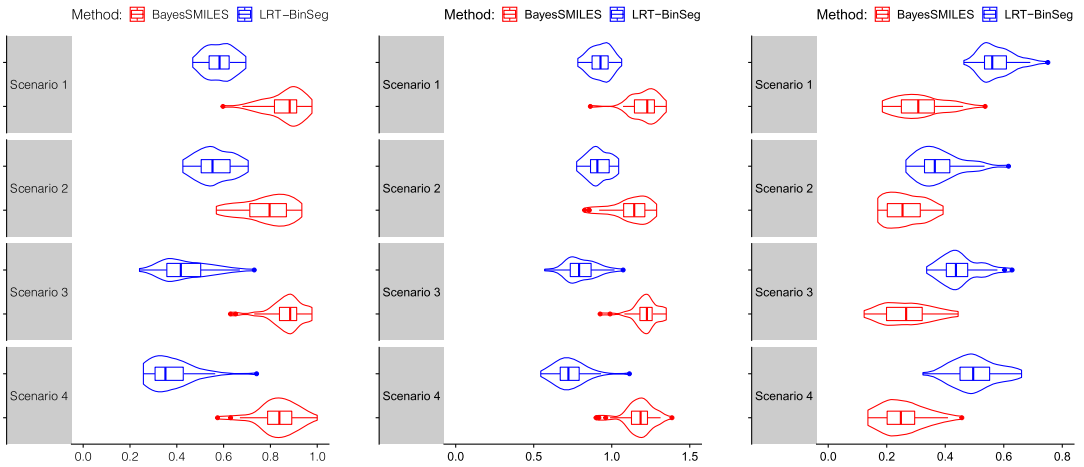


(b) Basic reproduction number estimation performance

Figure 3: Simulation study: The model fitting results based on a randomly selected simulated dataset (see the blue curve under Scenario 4 in Figure 2). (a) The locations of change points (blue solid lines) estimated from the posterior pairwise probability matrix (PPM) and their credible intervals (gray ribbons). The red dashed lines mark the true change point locations; (b) The posterior distributions of $\mathcal{R}_{0k}$'s for $k = 1, 2, 3, 4$ estimated from the segmented time-series data, given the three identified change points as shown in (a). The red dashed and blue solid lines are the true and estimated values of $\mathcal{R}_{0k}$'s, respectively. The two black solid lines are the lower and upper bounds of the 95% credible intervals.

setting up a comparison study, we therefore considered a two-stage approach that first identifies multiple change points of time-series data based on a likelihood based framework, and then estimates the basic reproduction numbers between each pair of nearby change points, following the stochastic SIR model introduced in Section 3.4. The alternative change point model assumes time points within one segment follow a normal distribution with distinct mean and/or variance from its nearby segments (Hinkley, 1970; Jen and Gupta, 1987), and it uses the likelihood ratio test (LRT) to detect multiple change points. An algorithm named binary segmentation (Edwards and Cavalli-Sforza, 1965; Sen and Srivastava, 1975) is commonly used to compute the test statistics for the LRT with high efficiency (Killick et al., 2012). In our case, to detect change points using this alternative approach named the likelihood ratio test with binary segmentation (LRT-BinSeg), we input the logarithmic scale of $\boldsymbol{I}$ into the function `cpt.meanvar` in the related `R` package `changepoint` (Killick and Eckley, 2014) for each of the simulated datasets. We set the maximum number of possible change points to 5 for the binary segmentation algorithm. Note that this restriction was not applicable to the alternative algorithms provided in the `changepoint` package. We also found that it tended to over-select the number of change points.

Figure 4(a) and (b) exhibit the change point detection performances for the four scenarios of $\boldsymbol{\mathcal{R}}_0$. Our BayesSMILES performed much better than the LRT-BinSeg with respect to change point detection under both performance metrics, ARI and MI. For instance, the ARI by BayesSMILES increased 39.29% to 122.16% over the LRT-BinSeg among the four scenarios, while the growth in MI could be up to 60.54%. Figure 4(c) compares the ability to capture the transmission dynamics in terms of RMSE, which depends on the change point detection accuracy. As expected, our BayesSMILES yielded smaller RMSE values across all scenarios since its identified change point locations were more accurate.



(a) Adjusted Rand index (ARI) **(b)** Mutual information (MI) **(c)** Root mean square error (RMSE)

Figure 4: Simulation study: The violin plots of (a) adjusted Rand index, (b) mutual information, and (c) $\boldsymbol{\mathcal{R}}_0$ root mean square error from 50 replicated datasets generated under the four scenarios. Red and blue violins correspond to the results obtained by BayesSMILES and LRT-BinSeg.

6 Analysis of COVID-19 Data

In this section, we applied BayesSMILES to the U.S. state-level COVID-19 daily report data provided by JHU-CSSE COVID-19 Data Repository.[1] Several recent COVID-19 studies also based their analyses on this resource (see e.g. Dong et al., 2020; Zhou and Ji, 2020; Toda, 2020). We first performed a preprocessing step to ensure the quality of the infectious data $\boldsymbol{I}$ for the model fitting. Due to the fact that recovery cases are not recorded in some states, we treated $\boldsymbol{I}$ and $\boldsymbol{R}$ as missing data and reconstructed the two sequences according to the process described in Section 3.1. The cumulative confirmed case numbers $\boldsymbol{C}$ were collected for each U.S. state starting from an early stage of the pandemic outbreak. In particular, we chose the starting time for each state as when there were at least ten confirmed COVID-19 cases for that state. We also set the removal rate $\gamma = 0.1$ as suggested by Pedersen and Meneghini (2020) and Weitz et al. (2020). Since different states could have different starting times, we further trimmed the sequences $\boldsymbol{I}$ and $\boldsymbol{R}$ for each state based on the latest starting time available. Finally, we set March 22, 2020, as the new starting time ($t = 1$) for all 50 states, and let July 19, 2020, be the last observed time point ($t = 120$).

We used the same hyperparameter and algorithm settings as described in Section 5.3. We ran four MCMC samplers for 40,000 iterations with the first half discarded as burn-in for the change point detection model to ensure reliable results. We randomly initialized the starting points for each chain. We assessed the concordance between the four chains based on the Pearson correlation coefficients of the marginal posterior probability of inclusions (PPIs), $\pi(\zeta_t|\cdot) \approx \sum_{u=1}^{U} \delta(\zeta_t^{(u)} = 1)/U$. For our real data analysis in this paper, we obtained coefficient values ranging from 0.951 to 0.997, which indicated good concordance among the four MCMC chains. Concordance among the marginal PPIs was confirmed by looking at their scatter plots across each pair of MCMC chains. Furthermore, we also used the Gelman and Rubin's convergence diagnostics (Gelman et al., 1992) to assess the convergence of the segment-specified basic reproduction numbers $\mathcal{R}_{0k}$'s to their posterior distributions. The potential scale reduction factors were all below 1.1, ranging from 1.001 to 1.045, clearly indicating that the MCMC chains for the stochastic SIR model were run for a satisfactory number of iterations, which was set to 100,000. Convergence was also confirmed by looking at their trace plots.

6.1 Detecting Change Points for U.S. States

We limit our analysis to four U.S. states with the highest cumulative confirmed cases as of July 19, 2020, to keep the paper in a reasonable length. They are New York, Texas, California, and Florida. The results for the 46 remaining states are available in https://shuangj00.github.io/BayesSMILES/ (see details in Supplement A). Figure 5 displays the detected change points, as well as the estimated basic reproduction number $\mathcal{R}_0$'s cross segments, for the four states. The associated credible interval to each identified change point is represented by a gray ribbon. In general, those change points detected by BayesSMILES indeed captured the important COVID-19 events that might affect the transmission rates. For instance, some change points reflected the positive effects of the preventative strategies such as lockdown, while others explained the "bounce back" in confirmed cases after reopening. Table S3 in the Supplementary Material lists the change point locations and their potentially related events for the four states.

[1] https://github.com/CSSEGISandData/COVID-19

In New York, the first change point was estimated to be March 28. We estimated the posterior mean of the basic reproduction number decreased from 2.24 (between March and March 27) to 1.63 (between March 28 and April 8). Notably, March 28 was the date when the Centers for Disease Control and Prevention (CDC) issued a 14-day domestic travel advisory for non-essential persons, which presumably alleviated the situation for the populated states such as New York. The second change point appeared around April 9, and the $\mathcal{R}_0$ of the third segment dropped to 0.98 with a 95% credible interval of [0.76, 1.25]. This matched the exact day when New York state posted its first drop in the ICU admissions since the COVID-19 outbreak began. The third change point was around April 27. Though there was no direct intervention issued in late April, we noticed that the mayor of New York City announced that all major events had been canceled starting from April 20. This action could bring a positive effect in controlling the outbreak, and our estimation from the SIR model suggested a further decrease in the basic reproduction number down to 0.66 with a 95% credible interval of [0.54, 0.81]. We observed another change point around June 18, which was close to the Phase II reopening of New York state on June 22. During Phase II reopening, restaurants were allowed to open for outdoor dining, stores opened for in-person retail, and more services resumed operational under strict limitations. Thus, we saw a little "bounced back" in $\mathcal{R}_0$ from 0.66 to 0.82. The last change point was on June 29. As expected, the basic reproduction number increased to 1.04 with a 95% credible interval of [0.84, 1.29] in the last segment. Although there was no public announcement around June 29 with a credible interval from June 28 to July 4, we suspect that the increased social interaction during the Independence Day long weekend (between July 3 and July 5) could be responsible for the increase in transmission dynamics.

In Texas, there were five change points detected. The first change point was estimated to be March 28, the same day as the first one for New York state. Due to a similar reason, the policy of mandatory 14-day quarantines for travelers entering Texas could bring a decrease in terms of the basic reproduction number (decreased from 2.97 to 2.07). The second change point was around April 9 with a further drop of $\mathcal{R}_0$ to 1.14 with a 95% credible interval [0.96, 1.35]. We found that the Texas Governor had extended the state's disaster declaration for an additional 30 days on April 12. The extension aimed at protecting the health and safety of Texans by ensuring adequate capabilities of supporting communities. Organizations such as the State Operations Center and the Strategic National Stockpile would continuously supply the state government with the resources needed to protect residents. May 25 was detected as the third change point, and it was the first time that $\mathcal{R}_0$ increased after the two drops. The estimated basic reproduction number was 1.29 with a 95% credible interval [1.02, 1.62]. This increase appeared around May 25 could be due to the Governor's updated executive order issued on May 26 that allowed additional services and activities to open for phase II reopening. The next change point was around June 16, and $\mathcal{R}_0$ further increased to 1.72 with a 95% credible interval [1.40, 2.11]. According to the prediction reported by the University of Texas at Austin's COVID-19 Modeling Consortium at the end of May, there might be a significant increase in the number of cases and hospitalizations beginning mid-June (News from *kxan*). Here, the change point location and the increased basic reproduction number were consistent with the results of this report. The last change point was around June 28 with an estimated decrease in $\mathcal{R}_0$ to 1.42 with a 95% credible interval [1.17, 1.72]. Notably, the Texas Governor issued multiple executive orders around late June to early July to mitigate the disease spreading. For instance, the executive order on June 26 reemphasized the limited occupancy for all business establishments in Texas. According to an executive order on July 2, all Texans were required to wear a face-covering in public spaces in

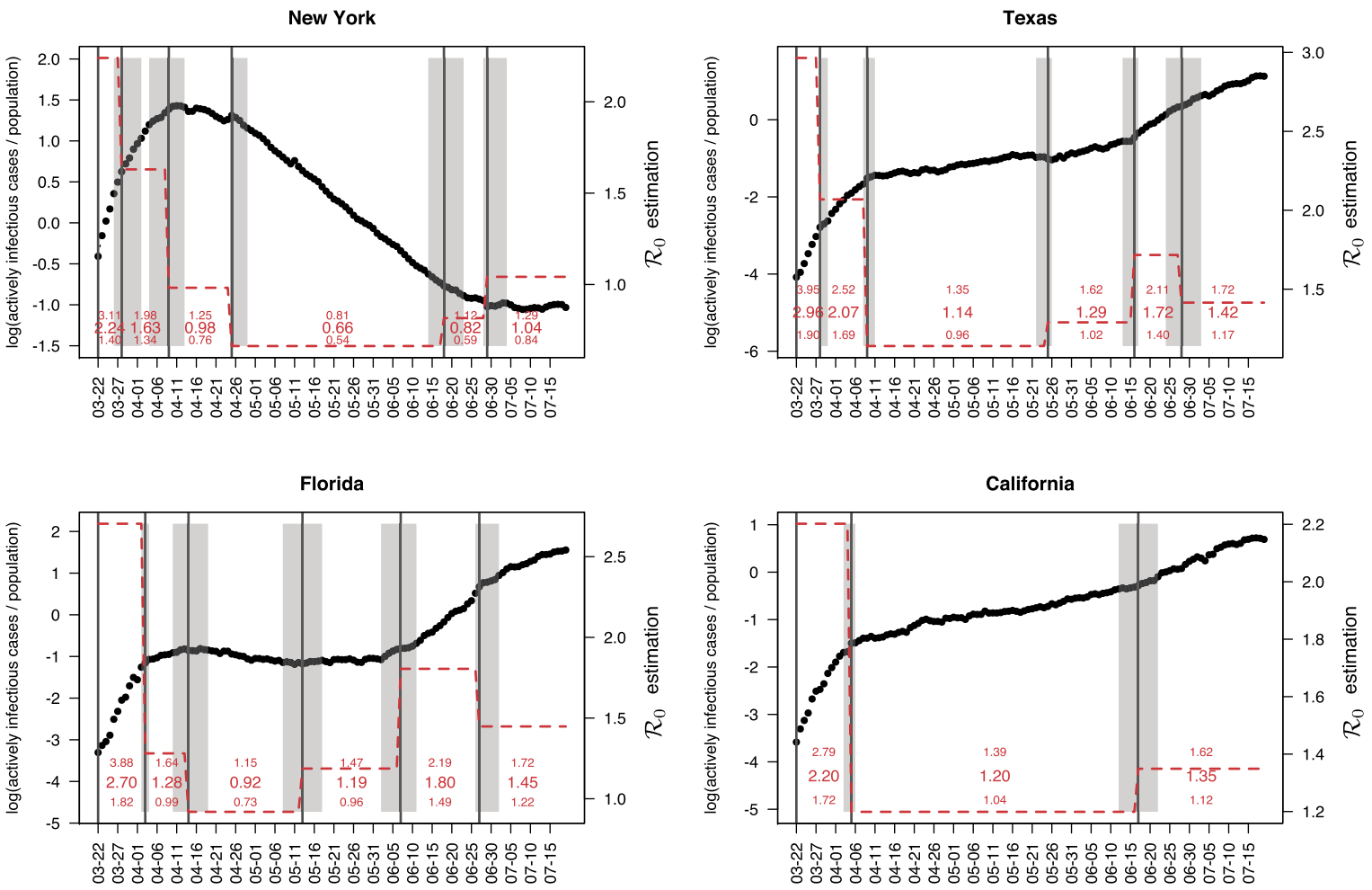


Figure 5: Case study: The change point detection and basic reproduction number estimation for the states of New York, Texas, Florida, and California. The black circles are the actively infectious case numbers divided by the total population (in thousands) at logarithmic scale, i.e. $log(I_t/N)$. The black solid lines pinpoint the change point locations, with the associated gray ribbons indicating the credible intervals. The red dashed lines describe the variation in the basic reproduction numbers $\boldsymbol{\mathcal{R}}_0$ across segments. The posterior means and the 95% credible intervals for $\mathcal{R}_{0k}$'s are given by red numbers.

counties with 20 or more positive COVID-19 cases. On the same day, the Governor announced an update regarding the executive order on June 26 with additional measures to slow the spread of COVID-19.

In Florida, the first estimated change point was April 3. It was two days after the statewide stay-at-home order for Florida. We estimated that the basic reproduction number decreased from 2.70 to 1.28 after the change point. The second change point appeared around the middle of April. Starting from April 13, some counties such as Osceola county enforced face-covering in public places. It could explain the reason why we observed a slight decrease in $\mathcal{R}_0$, from 1.28 to 0.92 with a 95% credible interval $[0.73, 1.15]$. The next change point was located around May 13, and $\mathcal{R}_0$ in this new stage went above 1 again, with a posterior mean of 1.19 and a 95% credible interval $[0.96, 1.50]$. We noticed that Florida entered the phase I reopening on May 18, which could lead to the "bounced back" situation. The fourth change point was around June 7, two days after the phase II reopening in Florida. Changes in the phase II reopening included that Universal Orlando opened the parks to the general public for the first time in months, and we observed that $\mathcal{R}_0$ increased again to 1.81 with a 95% credible interval $[1.50, 2.19]$. In the last segment (after June 27), our result revealed a slight drop in the basic reproduction number from 1.81 to 1.45. This change was potentially related to the consequence of requiring facial coverings in the four most populated cities in Florida: Tampa, Orlando, Miami, and Jacksonville. The face mask mandates went into effect for the four cities starting from June 19, 20, 25, and 29, respectively. Therefore, the drop in the transmissibility at the end of June may be explained by the effectiveness of wearing face masks as a non-pharmaceutical practice.

In California, we detected two change points. California was the first state to announce lockdown in the COVID-19 pandemic and its stay-at-home order became effective on March 19. Our change point detection results could miss these early actions since the data we analyzed started from March 22. The first selected change point was on April 5, with the value of the basic reproduction number decreasing dramatically when transitioning to the second segment (from 2.20 to 1.20). The second change point was on June 17, and we saw that $\mathcal{R}_0$ increased to 1.35 in the last segment with a 95% credible interval $[1.12, 1.62]$. According to California Governor, higher-risk businesses and venues (e.g. movie theaters, bars, gyms) were allowed to reopen with restrictions on June 12. Hence, the increase in the basic reproduction number could be the consequence of reopening. The same observation was made in New York and Texas.

6.2 Clustering U.S. States Based on Their Change Point Locations

We applied BayesSMILES on all 50 U.S. states. Based on the results, we seek to derive an overall picture of the COVID-19 dynamics across states. We summarized the temporally detected change points of the 50 states into common patterns, and then we labeled each state by matching its specific change point pattern to the common patterns. In particular, for each state, we calculated the marginal posterior probability of inclusion (PPI) for all time points, where the PPI for a time point t was calculated based on the B of MCMC samples after burn-in: $p_t^{\text{PPI}} = \sum_{b=1}^{B} \zeta_t / B$. Then we obtained the vector $\boldsymbol{p}^{\text{PPI}} = (p_1^{\text{PPI}}, \ldots, p_T^{\text{PPI}})$. Each entry in $\boldsymbol{p}^{\text{PPI}}$ is a value between 0 and 1, representing the proportion of time t selected as a change point among all iterations. Next, we computed the overall pattern by averaging over the vector $\boldsymbol{p}^{\text{PPI}}$ across 50 states. We noticed that some time points were rarely or never selected as change points. This naturally suggested that we could group the time points. To illustrate this, we trimmed the top 20% values of p_t^{PPI}

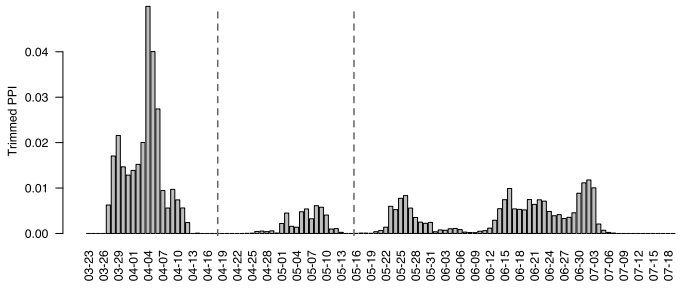


Figure 6: Case study: The averaged marginal posterior probability of inclusion (PPI) for each time point to be selected as a change point over all 50 U.S. states, after trimming the top 20% PPI values. The black dashed lines partition the whole time range into three segments: March 27–April 11, May 1–May 10, and May 22–July 3.

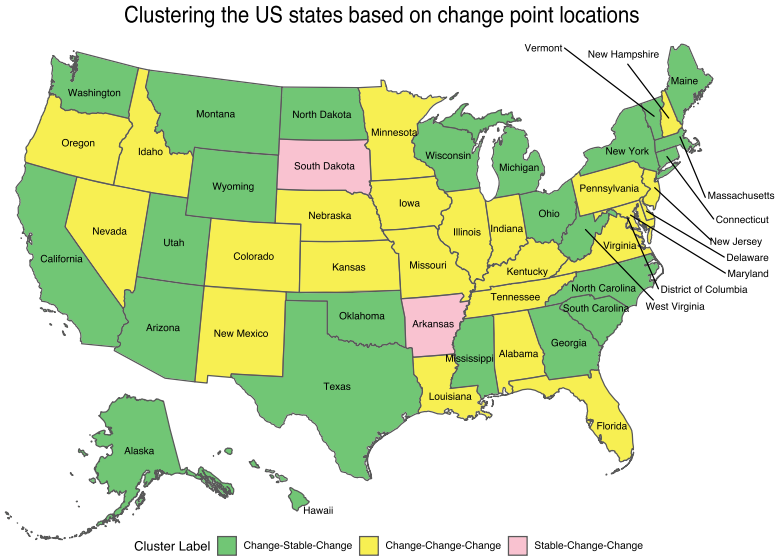


Figure 7: Case study: The temporal patterns of the COVID-19 transmission dynamics based on change points across the 50 U.S. states. Green, yellow, and pink correspond to "Change-Stable-Change", "Change-Change-Change", and "Stable-Change-Change" patterns, respectively.

for each time t (Figure 6). The trimming step provided a clear pattern and highlighted the groups of dates that were commonly identified as change points. We observed three time spans as shown in Figure 6: March 27–April 11, May 1–May 10, and May 22–July 3. For each state, we defined its cluster label based on the corresponding change point detection results. If a given state had at least one change point (including the credible interval) between March 27–April 11, the first element in its cluster label is "Change". Otherwise, the first element in the group label was set to "Stable". We repeated the same process to determine the second and third elements of the class label for each state. In the end, each state was assigned to a cluster label "Change-Change-Change", "Change-Stable-Change", or "Stable-Change-Change".

The map in Figure 7 colors each of the 50 states based on its cluster label, where green, yellow, and pink correspond to temporal patterns "Change-Stable-Change", "Change-Change-Change", and "Stable-Change-Change", respectively. Interestingly, three out of the four states

we analyzed, New York, Texas, and California, belonged to the same category, "Change-Stable-Change". Other states also in this category include Georgia, Arizona, North Carolina, and Louisiana. All of these states were in the top ten states with the most COVID-19 confirmed cases. We noticed that the phase I statewide reopening for all these states occurred in mid-May (May 15 for Georgia, May 13 for Arizona, May 8 for North Carolina, May 15 for Louisiana). Therefore, our model did not report any change points for these states between May 1 and May 10. The rest of the 10 states with the most COVID-19 confirmed cases, including Florida, Illinois, and New Jersey, were labeled as "Change-Change-Change", and all of them had a change point between May 1 and May 10. As discussed in Section 6.1, Florida had a change point around May 13 with a credible interval [May 8, May 18]. According to the Executive Order 2020-32 issued by the Illinois governor, the state entered the phase II reopening starting on May 1 with a modified stay-at-home order. For New Jersey, the statewide state-at-home order was not lifted until June 9. However, our model suggested a change point around the end of April with a credible interval [April 25, May 3] with a drop in $\mathcal{R}_0$ from 1.11 to 0.68 (details available at https://shuangj00.github.io/BayesSMILES/). We noticed that on May 3 the New Jersey governor announced a multi-state agreement to develop a regional supply chain for personal protective equipment, other medical equipment and testing. This joint-state protective measure allowed for efficient delivery and reliability of medical equipment for states and therefore best utilized life-saving resources in the face of the COVID-19 outbreak.

6.3 Predicting New Confirmed Cases for U.S. States

Reliable and accurate short-term forecasting of the new daily confirmed COVID-19 cases $\dot{C}_{T_f}$ at a future time T_f is important for both policy-makers and healthcare providers. We have illustrated how to use BayesSMILES to predict the new confirmed cases in Section 4.4. The idea is to make the short-term forecast based only on the observed data in the last available segment, ensuring that only the most recent disease characteristics are utilized. We compared BayesSMILES with the standard stochastic SIR model where all observed data from the first time point are used. We named this model FullDataSIR.

Figure 8 shows the true values of the new daily confirmed COVID-19 cases and the predictions made by BayesSMILES and FullDataSIR for the four major states. The 7-day forecast was chosen from July 20 to 26. First of all, it is observed that the predictive mean by FullDataSIR tended to be larger than that from BayesSMILES. This was because the basic reproduction numbers in the early stage (i.e. from late March to early April) were usually very large due to the lack of effective interventions. As a consequence of including those data, FullDataSIR inflated the predictions. We then quantified the prediction accuracy using the mean absolute percentage error (MAPE). The MAPE for the 7-day forecast is defined as

$$\text{MAPE}(\dot{\mathbf{C}},\ \hat{\dot{\mathbf{C}}}) \quad = \quad \frac{1}{7}\sum_{T_f=1}^{7}\left|\frac{\dot{C}_{T_f}-\hat{\dot{C}}_{T_f}}{\dot{C}_{T_f}}\right|,$$

where $\dot{C}_{T_f}$ and $\hat{\dot{C}}_t$ are the observed and predicted new confirmed cases at a future time T_f. The smaller the MAPE value, the more accurate the prediction. The numerical summary is shown in Table 1. For New York, Texas, and Florida, the MAPEs from BayesSMILES were much smaller than those from FullDataSIR, suggesting a better performance of BayesSMILES. However, for California, the two methods were almost the same in terms of short-term forecast.

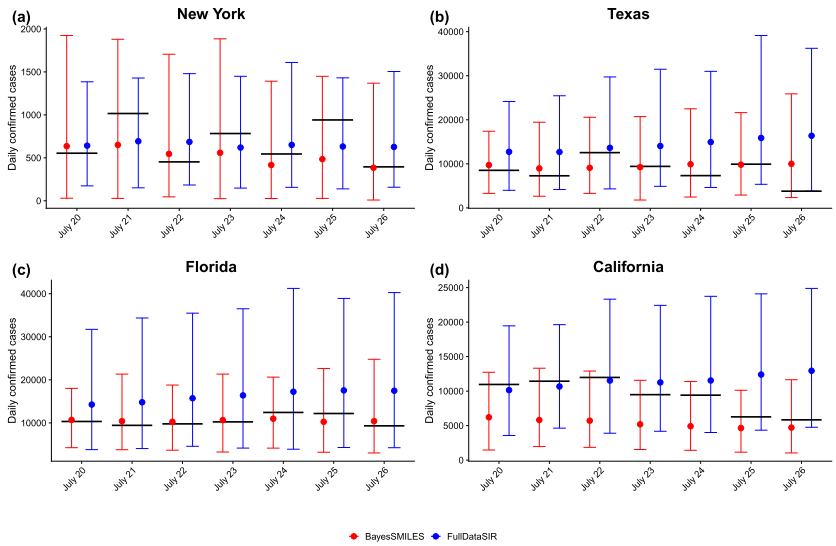


Figure 8: Case study: The 7-day forecast (between July 20 and 26) of the daily confirmed COVID-19 case numbers for the states of New York, Texas, Florida, and California. The red and blue circles and bars are the predictive means and 95% intervals by BayesSMILES and FullDataSIR, respectively. The black thick lines indicate the observed truth.

Table 1: Case study: The mean absolute percentage errors (MAPEs) of the 7-day forecast of daily confirmed COVID-19 case numbers by BayesSMILES and FullDataSIR for the states of New York, Texas, Florida, and California.

	New York	Texas	Florida	California
BayesSMILES	24.9	38.0	8.9	40.4
FullDataSIR	33.0	96.7	55.2	39.9

7 Conclusion

In this paper, we proposed BayesSMILES, a Bayesian segmentation model for analyzing longitudinal epidemiological data, to characterize the transmission dynamics of an infectious disease such as COVID-19. Our approach includes a Bayesian Poisson segmented regression model to detect multiple change points from the sequence of actively daily infectious cases. Those identified change points correspond to latent events that significantly altered disease spreading rates, while the resulting segments are characterized by unique epidemiological patterns. We further describe the disease transmissibility for each segment by using a stochastic time-invariant SIR model, assuming that the transmission rate remains the same until the next change point. Our model outputs a series of the basic reproduction numbers $\mathcal{R}_0$'s over stages to track the changes in spreading rates during a pandemic.

We applied BayesSMILES to analyze the COVID-19 daily report data of 50 U.S. states. Our results showed that the COVID-19 outbreak declined substantially after implementing stringent interventions for several states, including New York, Texas, and Florida. Meanwhile, our identified change points matched well with the timelines of publicly announced intervention strategies.

The change in the basic reproduction numbers between two adjacent segments might be used to quantify the effectiveness of an intervention, which could help us understand the impact of different control measures. Several downstream analyses based on the BayesSMILES results were conducted. In particular, we clustered the temporal patterns of the 50 U.S. states based on their change point locations, which led to an interesting spatial pattern related to the COVID-19 dynamics. Lastly, we demonstrated that our method could also improve the short-term forecasting of the new daily confirmed cases.

A potential issue of BayesSMILES is that the change point locations, which are identified in the Poisson segmented regression model, are set to be fixed when estimating the basic reproduction numbers using the stochastic SIR model. Such a two-stage approach might underestimate the uncertainties in $\mathcal{R}_{0k}$'s. A diagnostic method named simulation-based calibration (SBC) (Talts et al., 2018) is available to assess if the model inference has properly quantified the uncertainty. Using SBC to evaluate the soundness of the current MCMC sampling methods could be a future exploration. Another potential extension of the current work is to utilize advanced versions of the MCMC or Hamiltonian Monte Carlo (HMC) algorithms, which could lead to more accurate, efficient, and reliable inferences. For example, the MH with delayed rejection (Mira et al., 2001), the combination of delayed rejection and adaptive Metropolis samplers (Haario et al., 2006), the multiple-try Metropolis (Liu et al., 2000; Martino, 2018), as well as the methods discussed in Liang et al. (2011). One may also extend the Poisson error structure in the change point detection model to a negative binomial distribution for modeling the over-dispersed count data. Furthermore, the current BayesSMILES framework can be generalized to characterize temporal patterns in other epidemiological data. To do so, the segmented regression model should not be restricted to countable outcomes. Due to the concern for data accuracy, the result provided by the proposed method must be interpreted with caution. For instance, the number of confirmed cases is largely dependent on the test capacity and the number of recovery cases may suffer from under-reporting issues. How to improve the statistical power and prediction accuracy under those circumstances is worth investigating.

Supplementary Material

Supplement A: Software. We provide software in the form of `R/C++` codes on GitHub https://github.com/shuangj00/BayesSMILES. We have designed a website https://shuangj00.github.io/BayesSMILES/ to summarize the inference results for the 50 U.S. states, as a supplement to Section 6. The website shows that 1) the detected change points for each U.S. state; and 2) the COVID-19 transmission dynamics based on the segment-varying basic reproduction numbers $\mathcal{R}_0$'s, including their posterior means and 95% credible intervals.

Supplement B: Supplementary document. The supplementary file encloses the detailed Markov chain Monte Carlo algorithms, additional simulation study and real data analysis results, and key notation tables.

Acknowledgement

The authors would like to thank Jessie Norris for proofreading the manuscript.

Funding

This work was supported by the University of Texas at Dallas (UT Dallas) Office of Research [UT Dallas Center for Disease Dynamics and Statistics] and partially supported by the National Institutes of Health [1R56HG011035, 5P30CA142543, 5R01GM126479, 5R01HG008983].

References

Akiyama MJ, Spaulding AC, Rich JD (2020). Flattening the curve for incarcerated populations—COVID-19 in jails and prisons. *New England Journal of Medicine*, 382(22): 2075–2077.

Allen LJ (2008). An introduction to stochastic epidemic models. In: *Mathematical Epidemiology*, (F Brauer, P van den Driessche, J Wu, eds.), 81–130. Springer.

Alvarez FE, Argente D, Lippi F (2020). A simple planning problem for COVID-19 lockdown, *Technical report*, National Bureau of Economic Research.

Andersson H, Britton T (2012). *Stochastic Epidemic Models and Their Statistical Analysis*, volume 151. Springer Science & Business Media.

Bailey NT, et al. (1975). *The Mathematical Theory of Infectious Diseases and Its Applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE.

Becker NG, Britton T (1999). Statistical studies of infectious disease incidence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2): 287–307.

Chen YC, Lu PE, Chang CS (2020). A time-dependent SIR model for COVID-19. arXiv preprint: https://arxiv.org/abs/2003.00122.

Chowell G, Castillo-Chavez C, Fenimore PW, Kribs-Zaleta CM, Arriola L, Hyman JM (2004). Model parameters and outbreak control for SARS. *Emerging Infectious Diseases*, 10(7): 1258.

Cooper I, Mondal A, Antonopoulos CG (2020). A SIR model assumption for the spread of COVID-19 in different communities. *Chaos, Solitons & Fractals*, 139: 110057.

Dehning J, Zierenberg J, Spitzner FP, Wibral M, Neto JP, Wilczek M, et al. (2020). Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science*, 369(6500): eabb9789.

Dong E, Du H, Gardner L (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, 20(5): 533–534.

Edwards AW, Cavalli-Sforza LL (1965). A method for cluster analysis. *Biometrics*, 21(2): 362–375.

Flaxman S, Mishra S, Gandy A, Unwin HJT, Mellan TA, Coupland H, et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820): 257–261.

Gelman A, Lee D, Guo J (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5): 530–543.

Gelman A, Rubin DB, et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4): 457–472.

Giordano G, Blanchini F, Bruno R, Colaneri P, Di Filippo A, Di Matteo A, et al. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine*, 26: 855–860. 2020.

Gostic KM, McGough L, Baskerville EB, Abbott S, Joshi K, Tedijanto C, et al. (2020 Dec 10). Practical considerations for measuring the effective reproductive number, R_t. *PLoS Computational Biology*, 16(12): e1008409.

Haario H, Laine M, Mira A, Saksman E (2006). DRAM: efficient adaptive MCMC. *Statistics and Computing*, 16(4): 339–354.

Hinkley DV (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1): 1–17.

Hubert L, Arabie P (1985). Comparing partitions. *Journal of Classification*, 2(1): 193–218.

Jen T, Gupta AK (1987). On testing homogeneity of variances for Gaussian models. *Journal of Statistical Computation and Simulation*, 27(2): 155–173.

Kantner M, Koprucki T (2020). Beyond just "flattening the curve": Optimal control of epidemics with purely non-pharmaceutical interventions. *Journal of Mathematics in Industry*, 10(1): 1–23.

Kermack WO, McKendrick AG (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A*, 115(772): 700–721. Containing papers of a mathematical and physical character.

Killick R, Eckley I (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3): 1–19.

Killick R, Fearnhead P, Eckley IA (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500): 1590–1598.

Liang F, Liu C, Carroll R (2011). *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*, volume 714. John Wiley & Sons.

Liu JS, Liang F, Wong WH (2000). The multiple-try method and local optimization in Metropolis sampling. *Journal of the American Statistical Association*, 95(449): 121–134.

Lloyd-Smith JO, Galvani AP, Getz WM (2003). Curtailing transmission of severe acute respiratory syndrome within a community and its hospital. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270: 1979–1989. 1528.

Martino L (2018). A review of multiple try MCMC algorithms for signal processing. *Digital Signal Processing*, 75: 134–152.

Mira A, et al. (2001). On Metropolis-Hastings algorithms with delayed rejection. *Metron*, 59(3–4): 231–241.

Pedersen MG, Meneghini M (2020). Quantifying undetected COVID-19 cases and effects of containment measures in Italy. ResearchGate Preprint (online 21 March 2020). DOI: https://doi.org/10.13140/RG.2.2.11753.85600.

Rand WM (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336): 846–850.

Riley S, Fraser C, Donnelly CA, Ghani AC, Abu-Raddad LJ, Hedley AJ, et al. (2003). Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science*, 300(5627): 1961–1966.

Riou J, Hauser A, Counotte MJ, Althaus CL (2020). Adjusted age-specific case fatality ratio during the COVID-19 epidemic in Hubei, China, January and February 2020. medRxiv.

Roberts GO, Rosenthal JS (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1): 255–268.

Salvatier J, Wiecki TV, Fonnesbeck C (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2: e55.

Santos JM, Embrechts M (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification. In: *International Conference on Artificial Neural Networks*, (C Alippi, MM Polycarpou, CG Panayiotou, G Ellinas, eds.), 175–184. Springer.

Sen A, Srivastava MS (1975). On tests for detecting change in mean. *The Annals of Statistics*, 3(1): 98–108.

Steuer R, Kurths J, Daub CO, Weise J, Selbig J (2002). The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, 18(suppl 2): S231–S240.

Tadesse MG, Sha N, Vannucci M (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470): 602–617.

Talts S, Betancourt M, Simpson D, Vehtari A, Gelman A (2018). Validating bayesian inference algorithms with simulation-based calibration. arXiv preprint: https://arxiv.org/abs/1804.06788.

Toda AA (2020). Susceptible-infected-recovered (SIR) dynamics of COVID-19 and economic impact. arXiv preprint: https://arxiv.org/abs/2003.11221.

Vehtari A, Gelman A, Gabry J (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5): 1413–1432.

Wang L, Zhou Y, He J, Zhu B, Wang F, Tang L, et al. (2020 Jul 1). An epidemiological forecast model and software assessing interventions on the COVID-19 epidemic in China. *Journal of Data Science*, 18(3): 409–432.

Waqas M, Farooq M, Ahmad R, Ahmad A (2020). Analysis and prediction of COVID-19 pandemic in Pakistan using time-dependent SIR model. arXiv preprint: https://arxiv.org/abs/2005.02353.

Weitz JS, Beckett SJ, Coenen AR, Demory D, Dominguez-Mirazo M, Dushoff J, et al. (2020). Modeling shield immunity to reduce COVID-19 epidemic spread. *Nature Medicine*, 26: 849–854.

Zhou T, Ji Y (2020). Semiparametric Bayesian inference for the transmission dynamics of COVID-19 with a state-space model. arXiv preprint: https://arxiv.org/abs/2006.05581.